Data Analysis Formula Sheet

Fady Morris Ebeid
(2021)

Chapter 1 Practical Statistics

1 Descriptive Statistics

1.1 Data Types

- **Quantitative:** data takes on numeric values that allow us to perform mathematical operations (like the number of dogs).

  Quantitative data can be divided into:

  - **Continuous data:** can be split into smaller and smaller units, and still a smaller unit exists. An example of this is the age of the dog - we can measure the units of the age in years, months, days, hours, seconds, but there are still smaller units that could be associated with the age.
  - **Discrete data:** only takes on countable values. The number of dogs we interact with is an example of a discrete data type.

  They can also be classified into:

  - Interval data: numeric values where absolute differences are meaningful (addition and subtraction operations can be made). Examples: year and temperature in celsius.
  - Ratio data: numeric values where relative differences are meaningful (multiplication and division operations can be made). There must be a meaningful zero point. Examples: document word count and mass in kilograms.

- **Categorical (Qualitative):** are used to label a group or set of items (like dog breeds - Collies, Labs, Poodles, etc.).

  We can divide categorical data further into two types:

  - **Ordinal:** data take on a ranked ordering (like a ranked interaction on a scale from Very Poor to Very Good with the dogs). The distances between the categories are unknown.
  - **Nominal:** data do not have an order or ranking (like the breeds of the dog).

When analyzing categorical variables, we commonly just look at the count or percent of a group that falls into each level of a category.

1.2 Notation

A **random variable** is a placeholder for the possible values of some process. It represents a column in the dataset.
Random variables are represented by capital letters (for example X). Once we observe an outcome of these random variables, we notate it as a lower case of the same letter (for example x_1).

1.3 Summary Statistics

There are four main aspects to analyzing *quantitative* data.

1. Measures of Center
2. Measures of Spread
3. The Shape of the data
4. Outliers

Measures of Center

- **Mean:** (often called *average* or *expected value*)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  Where:
  $x_i \rightarrow$ The i^{th} data point.
  $n \rightarrow$ Number of samples.

- **Median:** The median splits our data so that 50% of our values are lower and 50% are higher.

$$\text{median}(X) = \begin{cases} x_{(n+1)/2}\,, & \text{if } n \text{ is odd} \\ \dfrac{x_{n/2} + x_{n/2+1}}{2}\,, & \text{if } n \text{ is even} \end{cases}$$

  In order to compute the median we must sort our values first.

- **Mode:** The mode is the most frequently observed value in our dataset.
  There might be multiple modes for a particular dataset(multimodal), or no mode at all.

Measures of Spread

The 5 number summary:

- **Minimum:** The smallest number in the dataset.
- **Q1 (First Quartile):** The value such that 25% of the data fall below.
- **Q2 (Second Quartile):** The value such that 50% of the data fall below (the median).
- **Q3 (Third Quartile):** The value such that 75% of the data fall below.
- **Maximum:** The largest value in the dataset.

Measures of Spread:

- **Range:** The difference between the *maximum* and the *minimum*.
- **Interquartile Range (IQR):** The difference between *Q3* and *Q1*.
- **Standard Deviation:** Is the average distance of each observation from the mean. Represents how far each point in our dataset is from the mean.

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  It has the same units as our original data.

- **Variance:** The average squared difference of each observation from the mean.

$$\text{Var}(X) = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

It has units that are the square of the units of the original data.

Shape of the Data

The shape of the data can be investigated using *histograms* or *box plots*.

The distribution of data can take one of three shapes:

- **Symmetric**: Normally distributed. The mean equals the median of the data.

$$\bar{x} = \text{median}(X)$$

  Examples: Height, Weight, Errors, Precipitation.

  To know if the data is normally distributed, there are plots called normal quantile-quantile plots and statistical methods like the Kolmogorov-Smirnov test.

- **Right skew** (positive skew - right tailed): The mean being skewed to the right of a median of the data.

$$\bar{x} > \text{median}(X)$$

  Examples: Amount of drug remaining in a blood stream, Time between phone calls at a call center, Time until light bulb dies.

- **Left skew** (negative skew - left tailed): The mean being skewed to the left of the median of the data.

$$\bar{x} < \text{median}(X)$$

  Examples: Grades as a percentage in many universities, Age of death, Asset price changes.

Outliers

Outliers are points that fall very far from the rest of our data points.
They influences measures like the mean and standard deviation much more than measures associated with the five number summary.
Outliers can be identified visually using histogram. And there are a number of different techniques for identifying outliers. Refer to [Seo06] paper.
When outliers are present we should consider the following points:

1. Noting they exist and the impact on summary statistics.
2. If typo - remove or fix
3. Understanding why they exist, and the impact on questions we are trying to answer about our data.
4. Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.
5. Be careful in reporting. Know how to ask the right questions.

1.4 General Steps for Working with a Random Variable

1. Plot your data to identify if you have outliers.
2. Handle outliers accordingly via the methods above.
3. If no outliers and your data follow a **normal** distribution - use the *mean* and *standard deviation* to describe your dataset, and report that the data are normally distributed.
4. If you have **skewed** data or outliers, use the *five number summary* to summarize your data and report the outliers.

1.5 Descriptive vs. Inferential Statistics

Comparison:

1. **Descriptive statistics** is about describing collected data.
2. **Inferential statistics** is about using collected data to draw conclusions to a larger population.
   We look at specific examples that allow us to identify the:
   (a) **population:** The entire group of interest.
   (b) **Parameter:** Numeric summary about the *population*.
   (c) **Sample:** Subset of the *population*.
   (d) **Statistic:** Numeric summary about a *sample*.

2 Probability

- Probability of an event: $P(A)$
- Probability of opposite event:
$$P(A) = 1 - P(\neg A)$$
- Probability of the occurance of a composite event n times (*independent*):
$$P(A, A, \ldots, A) = P(A) \cdot P(A) \cdot \ldots \cdot P(A) = P(A)^n$$

3 Binomial Distribution

The Binomial Distribution helps us determine the probability of x sucesses in n Bernoulli trials.
The probability mass function of the binomial distribution is given by:

$$b(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \ldots, n \\ 0, & \text{otherwise} \end{cases}$$

Where:
$p \rightarrow$ probability of success.
$x \rightarrow$ number of sucesses.
$n \rightarrow$ total number of trials.
$\binom{n}{x} \rightarrow$ binomial coefficient.

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Mean: $\mathbb{E}(X) = np$
Variance: $\text{Var}(X) = npq$

4 Conditional Probability

Conditional probability formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

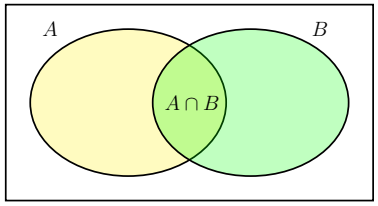
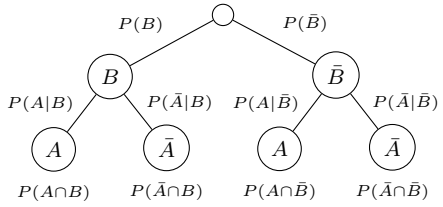


Figure 1.1: Conditional Probability

$$P(A \cap B) = P(A, B) = P(B)P(A|B)$$

$P(A, B)$ is the *joint probability* between A and B



Note that:

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$
$$P(\bar{A}) = P(\bar{A} \cap B) + P(\bar{A} \cap \bar{B})$$
$$P(A|B) + P(\bar{A}|B) = 1$$
$$P(A|\bar{B}) + P(\bar{A}|\bar{B}) = 1$$
$$P(A) + P(\bar{A}) = 1$$

5 Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B)$ is the *posterior probability*.
$P(A)$ is the *prior probability*.

5.1 Example: Cancer Test Case

Prior probabilities:

$$\begin{bmatrix} P(C) \\ P(\neg C) \end{bmatrix} = \begin{bmatrix} 0.01 \\ 0.99 \end{bmatrix}$$

Confusion Matrix:

$$\begin{array}{c} \\ C \\ \neg C \end{array} \begin{array}{c} \begin{array}{cc} Pos & Neg \end{array} \\ \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} \end{array} = \begin{bmatrix} P(Pos|C) & P(Neg|C) \\ P(Pos|\neg C) & P(Neg|\neg C) \end{bmatrix} = \begin{bmatrix} 0.9 & 0.1 \\ 0.15 & 0.85 \end{bmatrix}$$

Sensitivity (True Positive rate): measures the proportion of positives that are correctly identified (i.e. the proportion of those who have some condition (affected) who are correctly identified as having the condition). Also called the *recall*.
Specificity (True Negative rate): measures the proportion of negatives that are correctly identified (i.e. the proportion of those who do not have the condition (unaffected) who are correctly identified as not having the condition).
Type I error (false positive): "The true fact is that the patients do not have a specific disease but the physicians judges the patients was ill according to the test reports.d"
Type II error (false negative): "The true fact is that the disease is actually present but the test reports provide a falsely reassuring message to patients and physicians that the disease is absent."
Joint Probabilities (Interesection):

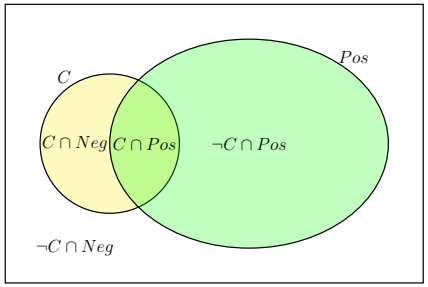


Figure 1.2: Cancer Test Case - Joint Probabilities

$$
\begin{aligned}
&\begin{bmatrix} P(Pos,C) & P(Neg,C) \\ P(Pos,\neg C) & P(Neg,\neg C) \end{bmatrix} \\
&= \begin{bmatrix} P(C) \\ P(\neg C) \end{bmatrix} \odot \begin{bmatrix} P(Pos|C) & P(Neg|C) \\ P(Pos|\neg C) & P(Neg|\neg C) \end{bmatrix} \\
&= \begin{bmatrix} P(C)P(Pos|C) & P(C)P(Neg|C) \\ P(\neg C)P(Pos|\neg C) & P(\neg C)P(Neg|\neg C) \end{bmatrix} \\
&= \begin{bmatrix} 0.009 & 0.001 \\ 0.1485 & 0.8415 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&\begin{bmatrix} P(Pos) & P(Neg) \end{bmatrix} \\
&= \begin{bmatrix} P(Pos,C) + P(Pos,\neg C) & P(Neg,C) + P(Neg,\neg C) \end{bmatrix} \\
&= \begin{bmatrix} 0.1575 & 0.8425 \end{bmatrix}
\end{aligned}
$$

Posterior probabilities:

$$
\begin{aligned}
&\begin{bmatrix} P(C|Pos) & P(C|Neg) \\ P(\neg C|Pos) & P(\neg C|Neg) \end{bmatrix} \\
&= \begin{bmatrix} P(Pos,C) & P(Neg,C) \\ P(Pos,\neg C) & P(Neg,\neg C) \end{bmatrix} \oslash \begin{bmatrix} P(Pos) & P(Neg) \end{bmatrix} \\
&= \begin{bmatrix} \dfrac{P(Pos,C)}{P(Pos)} & \dfrac{P(Neg,C)}{P(Neg)} \\ \dfrac{P(Pos,\neg C)}{P(Pos)} & \dfrac{P(Neg,\neg C)}{P(Neg)} \end{bmatrix} \\
&= \begin{bmatrix} 0.0571 & 0.0012 \\ 0.9429 & 0.9988 \end{bmatrix}
\end{aligned}
$$

$P(C|Pos)$ is called the *precision*.

6 Normal Distribution

$$
\begin{aligned}
\mathcal{N}\left(x;\mu,\sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}\right)} \\
&= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}
\end{aligned}
$$

7 Central Limit Theorem

Modeling coin flip according to number of flips:

Single Coin Flip	A few Coin Flips	Infinite coin flips (∞)
	Binomial distribution	Normal distribution
p	$\binom{n}{k} p^k (1-p)^{n-k}$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

7.1 Sampling Distribution

Definition 7.1. A **sampling distribution** is the distribution of a *statistic*. It is a distribution formed by samples.

Characteristics of a sampling distribution:

1. The sampling distribution is centered on the original parameter value.
$$\mu_M = \mu$$
2. The sampling distribution's variance depends on the sample size n. It decreases when n increases.
If we have a random variable $\mathbf{X}$, with a variance of σ^2, then the sampling distribution of the sample mean $\bar{\mathbf{X}}$ has a variance of
$$\sigma_M^2 = \frac{\sigma^2}{n}$$
3. The *standard error* is the *standard deviation* of the sampling distribution.
$$\text{standard error} = \sqrt{\frac{\sigma^2}{n}}$$

7.2 Notation

A **parameter** θ pertains to a *population*, while a **statistic** or **estimator** $\hat{\theta}$ pertains to a *sample*.

θ	Statistic		Description
μ	$\bar{x}$	$\hat{\mu}$	The mean of a dataset
σ	s	$\hat{\sigma}$	The standard deviation of a dataset
σ^2	s^2	$\hat{\sigma}^2$	The variance of a dataset
π	p	$\hat{\pi}$	The proportion (mean) of a binomial dataset
ρ	r	$\hat{\rho}$	The correlation coefficient
β	b	$\hat{\beta}$	The regression coefficient

A *binomial dataset* is a dataset with only 0 and 1 values.
The *parameter*, which is a numeric summary of the population doesn't change. While a *statistic* changes based on the sample selected from the population.

7.3 The Law of Large Numbers

Theorem 7.1. The **Law of Large Numbers**: As our sample size increases, the sample statistic gets closer to the population parameter.

Most common ways of parameter estimation:

1. Maximum Likelihood Estimation
2. Method of Moments Estimation
3. Bayesian Estimation

7.4 The Central Limit Theorem

Theorem 7.2. The **Central Limit Theorem** states that with a large enough sample size the sampling distribution of the mean will be normally distributed.

It applies for the following statistics:

1. Sample means ($\bar{x}$).
2. Sample proportions(p).
3. Difference in sample means ($\bar{x}_1 - \bar{x}_2$).
4. Difference in sample proportions ($p_1 - p_2$).

It does *not* apply to the following statistics:

1. Sample standard deviation s.
2. Correlation coefficient r.
3. Maximum value in the dataset.

7.5 Bootstrapping

Definition 7.2. Bootstrapping: is a technique where we sample from a group *with replacement*.

- We can use bootstrapping to simulate the creation of sampling distribution.
- An element can be picked more than once from the dataset.
- The probability of any number in our set stays the same regardless of how many times it has been chosen. Flipping a coin and rolling a die are examples of bootstrapping.

8 Confidence Intervals

Definition 8.1. Confidence intervals provide a range of values that are possible for a population *parameter*. A confidence interval is the probability that a population parameter will fall between a set of values for a certain proportion of times.

- **Confidence intervals** can be interpreted as "we are x% confident that the *population* parameter falls between the bounds of the interval"
- Confidence intervals can be built for different parameters such as population mean, or difference in means.
- Confidence levels can be 90%, 95%, 98%, 99%
- An important application that uses comparison of means is A/B testing.

8.1 Statistical vs. Practical Significance

Statistical significance: Evidence from hypothesis tests and confidence intervals that H_1 is true.
Practical significance: Considers real world aspects, not just numbers in making final conclusions. It takes into account other real world constraints such as *space*, *time*, or *money*.

8.2 Building Confidence Intervals

There are two methods:

- Bootstrapping.
- Traditional Methods: These methods are no longer necessary with what is possible with statistics in modern computing. For reference, see Stat Trek site

8.3 Other Terms Associated with Confidence Intervals

- The confidence interval **width**: is the difference between the upper and lower bounds of the confidence interval.
- The **margin of error**: Is half the confidence interval width.
  Example: "Candidate A has $34\% \pm 3\%$ of the votes" $\implies (31\%, 37\%)$

The relationship between *sample size* and *confidence level* to *confidence interval*:

1. Increasing the sample size n will decrease the width of confidence interval.
2. Increasing the confidence level (say from 95% to 99%) will increase the width of the confidence interval.

8.4 Confidence Intervals vs. Machine Learning

Confidence intervals are about parameters in a *population*, while machine learning make predictions about *individual data points*.

9 Hypothesis Testing

Rules of setting up the *null* and *alternative* hypotheses:

1. The H_0 is the *null hypothesis*. It is the condition we believe to be true before collecting any data.
2. The H_0 usually states that there is no effect or that two groups are equal.
3. H_1 is the *alternative hypothesis*. It is what we would like to prove to be true.
4. The H_0 and H_1 are competing, non-overlapping hypotheses.
5. H_0 contains an equal sign of some kind when it pertains to mathematical ideas. Either $=, \leq$, or $\geq$.
6. H_1 contains the opposition of the null hypothesis. Either $=, >$, or $<$.

9.1 Types of Errors

Table of error types		Null Hypothesis (H_0) is	
		False (+)	True (−)
Decision about null hypothesis (H_0)	Reject (+)	True Positive $1-\beta$	False Positive (Type I error) α
	Don't reject (−)	False Negative (Type II error) β	True Negative $1-\alpha$

Type I Errors:

- False positives.
- Error rate is denoted by α. Commonly it is 1-5%.
- Deciding the alternative (H_1) is true, when actually the null (H_0) is true
- The worse of errors.
- You should set up your null and alternative hypotheses, so that the worst of your errors is the type I error.

Type II Errors:

- False negatives.
- Error rate is denoted by β.
- Deciding the null (H_0) is true, when actually the alternative (H_1) is true.

Power of a statistical test: True positive rate $(1-\beta)$. This is the ability of an individual to correctly choose the alternative hypothesis (the probability of rejecting a null hypothesis that is false).

9.2 Common Types of Hypothesis Tests

Hypothesis tests are preformed on **population parameters**, never on statistics, as statistics are values you already know from the data.
Common hypothesis tests:

1. Testing a population mean (One sample t-test).
2. Testing the difference in means (Two sample t-test).
3. Testing the difference before and after some treatment on the same individual (Paired t-test).
4. Testing a population proportion (One sample z-test).
5. Testing the difference between population proportions (Two sample z-test).

T table: t distribution critical values.
Two-sided test: Testing simply if the parameters of two groups are the same or if they are different. The equal case should still be in the null hypothesis. We aren't interested in if one parameter is greater than another.

9.3 Difference in Means

Notice the standard deviation with the difference in means is actually the square root of the sum of the variance of each of the individual sampling distributions.

$$\sigma_{\text{diff}} = \sqrt{\sigma_{M1}^2 + \sigma_{M2}^2} \implies \sigma_{\text{diff}}^2 = \sigma_{M1}^2 + \sigma_{M2}^2$$

And the standard deviation of the mean is the standard deviation of the original draws divided by the square root of the sample size taken.

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

Thus:

$$\sigma_{\text{diff}}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

9.4 *P*-value

***P*-value**: The conditional probability of observing a test statistic (or more extreme test results) in favor of the alternative hypothesis (H_1) if the null hypothesis (H_0) is true.

- A very small p-value means that such an extreme observed outcome would be very unlikely under the null hypothesis.
  Small p-value $\longrightarrow$ Choose H_1
- It is actually always the case that when your p-value is large you will end up staying with the null hypothesis as your choice.
  Large p-value $\longrightarrow$ Choose H_0

One-Sided Right-Tail Test

Consider an observed test-statistic t from unknown distribution T. The p value is the prior probability of observing a test statistic $\geq t$ if H_0 were true.

The process is as follows:

1. Simulate the values of your statistic that are possible from the null.
2. Calculate the value of the statistic (t) you actually obtained in your data.
3. Compare your statistic to the values from the null.
4. Calculate the p-value as the proportion of null values that are considered extreme based on your alternative.

$$p = P(T \geq t | H_0 \text{ is true})$$

p-value and Errors

Compare p-value and type I error threshold (α), the decision about which hypothesis to choose becomes:

- p-value $\leq \alpha \implies$ Reject H_0
- p-value $\geq \alpha \implies$ Fail to Reject H_0

9.5 Other Things to Consider

Impact of Large Sample Size

- With large sample sizes, hypothesis testing leads to even the smallest findings become *statistically significant*(ending up rejecting essentially every null). However, these findings many not be practically significant at all.
- Hypothesis testing takes an aggregate approach towards the conclusions made based on data, as these tests are aimed at understanding population parameters (which are aggregate population values).
- Alternatively, machine learning techniques take an individual approach towards making conclusions, as they attempt to predict an outcome for each specific data point.

Performing More Than One Hypothesis Test

When performing more than one hypothesis test, your type I error compounds. In order to correct for this, you can use one of the following techniques:

1. **Bonferroni** correction.

   A simple, but very conservative approach.

   The new type I error rate should be the error rate you actually want divided by the number of tests you are performing.

   Let m be the number of tests, then your new error rate becomes:

   $$\alpha_{\text{new}} = \frac{\alpha}{m}$$

2. Turkey correction
3. Q-values (popular in medical tests).

How Do Confidence Intervals and Hypothesis Testing Compare

Confidence interval and two-sided hypothesis test (a test that involves a $\neq$ in the alternative) are the same in terms of conclusions.

$$1 - \text{CI} = \alpha$$

Example: 95% confidence interval is similar to a hypothesis test with a type I error reate $\alpha = 0.05$

---

10 A/B Testing

Definition 10.1 (A/B testing)**.** (also known as *bucket testing* or *split-run testing*) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

A/B testing is a form of hypothesis testing where:

- **Null Hypothesis:** Experiment does equally or worse than the control.
- **Alternative Hypothesis:** Experiment does better than the control.

A/B testing has drawbacks. It can't tell you about options you haven't considered. It is also subject to bias when tested on existing users. There are two types of bias:

- **Change Aversion:** Existing users may give an unfair advantage to the *old version*, simply because they resist change, even if it is better on the long run.
- **Novelty Effect:** Existing users may give an unfair advantage to the *new version*, because they are excited or drawn to change, even if it isn't any better in the long run.

10.1 Testing Changes on a Web Page

A/B tests are used to test changes on a web page by running an experiment where a **control group** sees the old version, while the **experiment group** sees the new version. A **metric** is then chosen to measure the level of engagement from users in each group. These results are then used to judge whether one version is more effective than the other.

The metric used is the **click through rate (CTR)**.

$$\textbf{CTR} = \frac{\textbf{\# clicks by unique users}}{\textbf{\# views by unique users}}$$

Hypotheses:

$$H_0 : \text{CTR}_{\text{new}} \leq \text{CTR}_{\text{old}}$$
$$H_1 : \text{CTR}_{\text{new}} > \text{CTR}_{\text{old}}$$

Steps Taken to Analyze the Results of A/B test

1. Compute the **observed difference**, between the CTR metrics for the control and experiment group.
2. Simulate the **sampling distribution** for the difference in proportions(difference in CTR).
3. Use the sampling distribution to simulate the **distribution under the null** hypothesis, by creating a random normal distribution centered at 0 with the same standard deviation as the sampling distribution.
4. Compute the ***p*-value** by finding the null values in the null distribution that are greater than the observed difference.
5. Compare the p-value to type I error threshold α to determine the **statistical significance** of observed difference.

Analyzing Multiple Metrics

The more metrics to evaluate, the more likely to observe significant just by chance. The probability of any false positive increases as you increase the number of metrics(tests). This multiple comparison problem can be solved by several techniques, such as bonferroni correction (see section 9.5).

Since the Bonferroni method is too conservative when we expect correlation among metrics, we can better approach this problem with more sophisticated methods, such as

- The closed testing procedure
- Boole-Bonferroni bound
- The Holm-Bonferroni method.

These are less conservative and take this correlation into account.

If you do choose to use a less conservative method, just make sure the assumptions of that method are truly met in your situation, and that you're not just trying to cheat on a p-value. Choosing a poorly suited test just to get significant results will only lead to misguided decisions that harm the performance in the long run.

Difficulties in A/B Testing

- Novelty effect and change aversion when existing users first experience a change
- Sufficient traffic and conversions to have significant and repeatable results
- Best metric choice for making the ultimate decision (eg. measuring revenue vs. clicks)
- Long enough run time for the experiment to account for changes in behavior based on time of day/week or seasonal events.
- Practical significance of a conversion rate (the cost of launching a new feature vs. the gain from the increase in conversion)
- Consistency among test subjects in the control and experiment group (imbalance in the population represented in each group can lead to situations like Simpson's Paradox)

---

11 Regression

Machine learning is split into:

- **Supervised learning:** Predicting the labels of data. Examples are fraud detection, whether customers will buy a product or not, house values.
- **Unsupervised learning:** Clustering unlabeled data together.

11.1 Simple Linear Regression

We compare two quantitative variables to one another. To predict a linear function.

$$y = f(x)$$

y is called the **response** variable or **dependent** variable. It is the variable you are interested in predicting.

x is called the **explanatory** variable or **independent** variable. It is the variable used to predict the response.

11.2 Covariance

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

11.3 Correlation Coefficient

There are different ways to measure the correlation between two variables [see this link].

For a *linear relationship*, the most common way to measure correlation is Pearson's correlation coefficient.

Denoted by r.

$$r \in [-1, 1]$$

Correlation coefficient for a sample:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

It measures:

- Strength: A relationship can be classified according to strength into:
  - **Weak:** $0.0 \leq |r| < 0.3$
  - **Moderate:** $0.3 \leq |r| < 0.7$
  - **Strong:** $0.7 \leq |r| \leq 1.0$
- Direction: Either negative or positive depending on the sign.

11.4 Equation of The Line

The line in *linear regression* has the equation:

$$\hat{y} = b_0 + b_1 x_1$$

Where:
$\hat{y} \longrightarrow$ is the **predicted** value of the response from the line.
$b_0 \longrightarrow$ is the **intercept**. It is the predicted value of the response when the x-variable is zero. (denoted β_0 for the population).
$b_1 \longrightarrow$ is the **slope**. It is the predicted change in the response for every one unit increase in the x-value. (denoted β_1 for the population).
$x_1 \longrightarrow$ is the **explanatory variable**.
$y \longrightarrow$ is the **actual** response value for a data point in our dataset(also called **label**).

11.5 Fitting a Regression Line

The algorithm used to fit a regression line to a dataset is called **least-squares** algorithm, it finds the line that minimizes

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Closed form solution:

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\bar{y} = \frac{1}{n}\sum y_i$$

$s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$ (Using the Bessel's Correction formula).

$s_y = \sqrt{\frac{1}{n-1}\sum (y_i - \bar{y})^2}$ (Using the Bessel's Correction formula).

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

$$b_1 = r\frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Chapter 2
Data Visualization

There are two main reasons for creating visuals using data:

1. **Exploratory** analysis is done when you are searching for insights. These visualizations don't need to be perfect. You are using plots to find insights, but they don't need to be aesthetically appealing. You are the consumer of these plots, and you need to be able to find the answer to your questions from these plots.
2. **Explanatory** analysis is done when you are providing your results for others. These visualizations need to provide you the emphasis necessary to convey your message. They should be accurate, insightful, and visually appealing.

The five steps of the data analysis process:

1. **Extract** - Obtain the data from a spreadsheet, SQL, the web, etc.
2. **Clean** - Here we could use exploratory visuals.
3. **Explore** - Here we use exploratory visuals.
4. **Analyze** - Here we might use either exploratory or explanatory visuals.
5. **Share** - Here is where explanatory visuals live.

1 Design of Visualizations

Visuals can be bad if they:

1. Don't convey the desired message.
2. Are misleading.

Reference [Huf93]

1.1 Visual Encodings

humans are able to best understand data encoded with:

- **positional changes** (differences in x- and y- position as in scatterplots)
- **length changes** (differences in box heights as bar charts and histograms).

Alternatively, humans struggle with understanding data encoded with:

- **color hue changes** (as are unfortunately commonly used as an additional variable encoding in scatter plots)
- **area changes** (as in pie charts, which often makes them not the best plot choice).

1.2 Chart Junk

Chart junk: all visual elements in charts and graphs that are not necessary to comprehend the information represented on the graph or that distract the viewer from this information.
Examples of chart junk include:

1. Heavy grid lines
2. Unnecessary text
3. Pictures surrounding the visual
4. Shading or 3d components
5. Ornamented chart axes

1.3 Data-Ink Ratio

The more of the ink in your visual that is related to conveying the message in the data, the better.
Limiting chart junk increases the data-ink ratio.

1.4 Design Intergrity

Lie factor: the degree to which visualization distorts or misinterprets the data values being plotted. It is calculated as:

$$\text{lie factor} = \frac{\Delta\text{visual}/\text{visual}_{\text{start}}}{\Delta\text{data}/\text{data}_{\text{start}}}$$

It is the relative change shown in the graphic divided by the actual relative change in the data.
Ideally, the lie factor should be 1. Any other value means that there is some mismatch in the ratio of depicted change to actual change.

1.5 Using Color

1. Before adding color to a visualization, start with black and white.
2. When using color, use less intense colors - not all the colors of the rainbow, which is the default in many software applications.
3. Color for communication. Use color to highlight your message and separate groups of interest. Don't add color just to have color in your visualization.

1.6 Designing for Color Blindness

Stay away from **red to green palette** and use **blue to orange palette** instead.
Further reading: 5 tips on designing colorblind-friendly visualizations

1.7 Additional Encodings

We typically use the $x-$ and $y-$ axes to depict the value of variables. If we have more than two variables we can use other visual encodings:

1. **color and shape** for **catigorical** data.
2. **size of marker** for **quantitative** data.

Use additional encoding only when necessary. If the visual gets complicated consider breaking it into multiple visuals that convey **multiple messages**.

2 Univariate Exploration of Data

Univariate visualizations: Visualize single-variables, such as bar charts, histograms, and line charts.

2.1 Bar Charts

Used to depict the distribution of categorical/qualitative variables. They can also be used for discrete quantitative data.

2.2 Pie Charts

Guidelines to use a pie chart:

- Make sure that your interest is *relative* frequencies.
- Limit the number of slices plotted to two or three slices, though it's possible to plot with four or five slices as long as the wedge sizes can be distinguished.
- Plot the data systemically. Start from 12 o'clock position of the circle, then plot each categorical level from most frequent to least frequent.

Otherwise, use a bar chart instead.

2.3 Histograms

A histogram is used to plot the distribution of a quantitative(numeric) variable. It is the quantitative version of the bar chart. Values are grouped into continuous bins.
When a data value is on a bin edge, it is counted in the bin to its right. The exception is the rightmost bin edge.

3 Bivariate Exploration of Data

Bivariate visualizations: Those visualizations involving two variables. The variation in one variable will affect the value of the other variable.

3.1 Scatterplots

Quantitative variable vs. quantitative variable.
If we have a very large number of points or our numeric variables are discrete-valued, then using a scatterplot won't be informative. The visualization will suffer from *overplotting*.
overplotting is where the high amount of overlap in points makes it difficult to see the actual relationship between variables.
To make the trends in the data clearer, it is overcome by:

1. Using jitter.
2. Using transparency.

3.2 Heat Maps

Quantitative variable vs. quantitative variable.
A **heat map** is a 2-D version of the histogram that can be used as an alternative to a scatterplot.
They are good in the following cases:

1. Good for discrete variable vs. discrete variable.
2. Good alternative to transparency for a lot of data.

Correct choice of bin sizes is important.

3.3 Violin Plots

Quantitative variable vs. qualitative(categorical) variable.
For each level of categorical variable, a distribution of the values on the numeric variabble is plotted.
The distribution is plotted as a kernel density estimate, like a smoothed histogram.

3.4 Box Plots

Quantitative(numeric) vs. qualitative(categorical).
Compared to the violin plot, the box plot is better for displaying the 5 point summary of data, reporting a set of descriptive statistics:

- The central line in the box indicates the median of the distribution
- The top and bottom of the box represent the third and first quartiles of the data, respectively.
- The height of the box is the interquartile range (IQR).
- From the top and bottom of the box, the whiskers indicate the range from the first or third quartiles to the minimum or maximum value in the distribution.
- Typically, a maximum range is set on whisker length; by default, this is 1.5 times the IQR.
- Individual points below the lower whisker or above the upper whisker indicate individual outlier points that are more than 1.5 times the IQR below the first quartile or above the third quartile.

Box plots are better than violin plots for *explanatory* visualizations.

3.5 Clustered Bar Charts

Depict the relationship between two categorical variables.
Bars are organized into clusters based on levels of the first variable, and then bars are ordered consistently across the second variable within each cluster.

3.6 Line Plots

Line plot: plots the trend of one numeric variable against values of a second variable. In a line plot, only one point is plotted for every unique x-value or bin of x-value (like a histogram).
Line plots are used instead of bar plots to:

- Emphasize relative change. The need for a zero on the y-axis is not necessary.
- Emphasize trends across x-values.

Time series plot: A line plot where the x-variable represents time. For example, stock or currency charts.

4 Multivariate Exploration of Data

4.1 Non-Positional Encodings for Third Variables

There are four major cases to consider when we want to plot three variables together:

- Three numeric variables
- Two numeric variables and one categorical variable
- One numeric variable and two categorical variables
- Three categorical variables

If we have at least two numeric variables, we use scatterplot to encode two of the numeric variables, then use a non-positional encoding (like shape, size and color) for the third variable.

4.2 Color Palettes

There are three major classes of color palette to consider:

- **Qualitative:** Distinct colors, for nominal-type data.
- **Sequential:** light-to-dark trend across a single or small range of hues of the same color. Used for categorical ordinal or numeric(quantitative) data types.
- **Diverging:** Used if there is a meaningful zero or center value for the variable.Two sequential palettes with different hues are put back to back, with common color (usually white or gray) connecting them. One hue indicates values greater than the center point, while the other indicates values smaller than the center.

4.3 Faceting

Faceting allows you to plot multiple simpler plots across levels of one or two other variables.

4.4 Heat Maps

They can be used for multivariate visualizations, substituting count with a third variable.

4.5 Plot Matrices

Give a high level look at pairwise relationship between all variables. In a plot matrix, each row and column represents a different variable.

4.6 Correlation Heat Maps

For numeric variables this is similar to a correlation matrix. It shows the strength of relationships between variables.

4.7 Feature Engineering

There may exist two variables that are related in some way. Feature engineering is about creating a new variable with a sum, product or ratio between original variables that give insight into research questions.
Example:

$$\text{crime incident rate} = \frac{\text{crime incidents}}{\text{population totals}}$$

Another way to perform feature engineering is to divide a numeric variable into ordered bins.

References

[Huf93] Darrell Huff. *How to Lie with Statistics*. English. Reissue edition. New York: W. W. Norton & Company, Oct. 1993. ISBN: 978-0-393-31072-6.

[Seo06] Songwon Seo. *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. en. University of Pittsburgh ETD. Publisher: University of Pittsburgh. Aug. 2006. URL: http://d-scholarship.pitt.edu/7948/ (visited on 01/28/2021).